**Name: Abraham. P**

**Date: 28-Mar-2025**

**Course: DA & DS**

**Batch: Feb 2025**

**Advanced SQL – Reinforcement Project – IMDB Dataset**

The dataset provided is a simplified version of the IMDb database, structured to capture essential information about movies, their genres, actors, directors, ratings, and more. This database consists of several tables that contain various details such as:

1. Movie: Contains basic information about each movie, including title, release year, duration, country, income, languages, and production companies.

2. Genre: Describes the genres associated with each movie.

3. Director Mapping: Maps movies to their directors.

4. Role Mapping: Maps actors/actresses to movies and specifies the role category (e.g., actor, director, producer).

5. Names: Stores information about people (actors, directors, etc.), including their birthdates, heights, and known movies.

6. Ratings: Contains ratings information for movies, including the average rating, total votes, and median rating.

Queries & Screenshot’s

1. Count the total number of records in each table of the database.

To retrieve the total number of records in each table, we use the **COUNT** aggregate function in the **SELECT** statement. The query below is used to obtain the total record count for each table in this database.

Here are the total record counts for each table in the database:

Total number of movies: 7,997

Total number of genres: 14,662

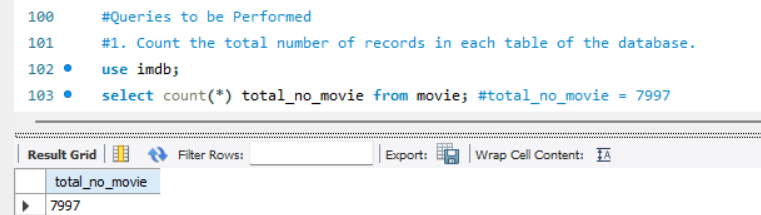
Total number of directors: 3,867

Total number of names: 25,735

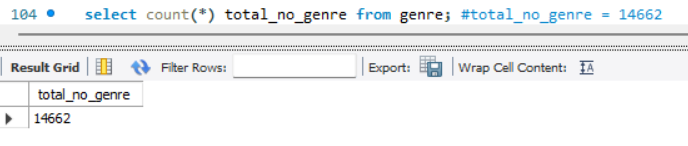
Total number of ratings: 7,997

Total number of role mappings: 15,615

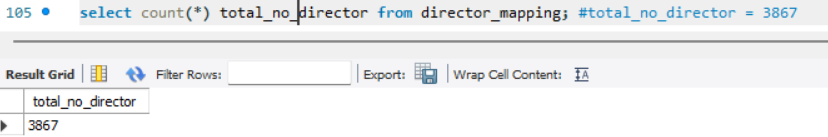
select count(\*) total\_no\_movie from movie; #total\_no\_movie = 7997



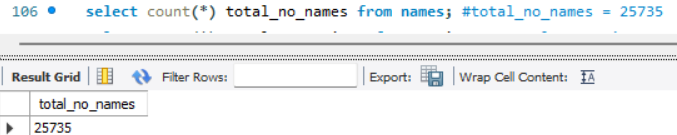
select count(\*) total\_no\_genre from genre; #total\_no\_genre = 14662



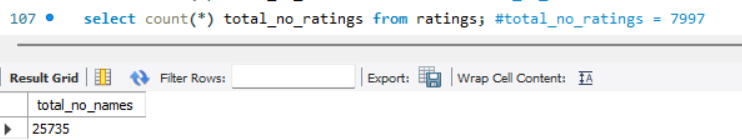
select count(\*) total\_no\_director from director\_mapping; #total\_no\_director = 3867



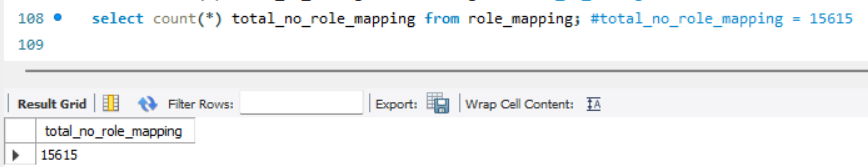
select count(\*) total\_no\_names from names; #total\_no\_names = 25735



select count(\*) total\_no\_ratings from ratings; #total\_no\_ratings = 7997



select count(\*) total\_no\_role\_mapping from role\_mapping; #total\_no\_role\_mapping = 15615

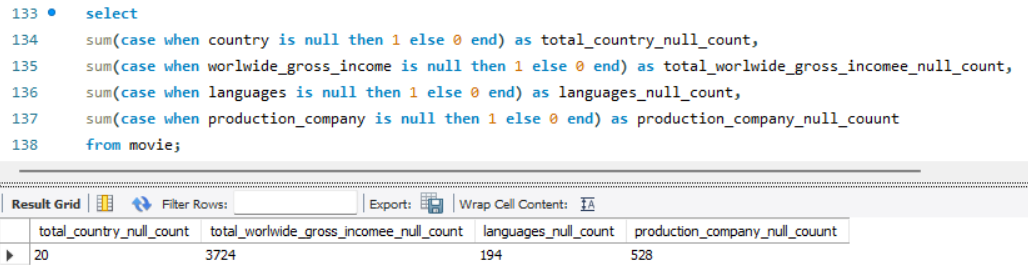


1. Identify which columns in the movie table contain null values.

Country, worldwide\_gross\_income, languages and production\_company column’s have null values.

To identify null values in a table, I use the IS NULL operator in the WHERE condition. In the movie table, I checked all columns and found the following null value counts:

* country column has 20 null rows.
* worldwide\_gross\_income column has 3,724 null rows.
* languages column has 194 null rows.
* production\_company column has 528 null rows.

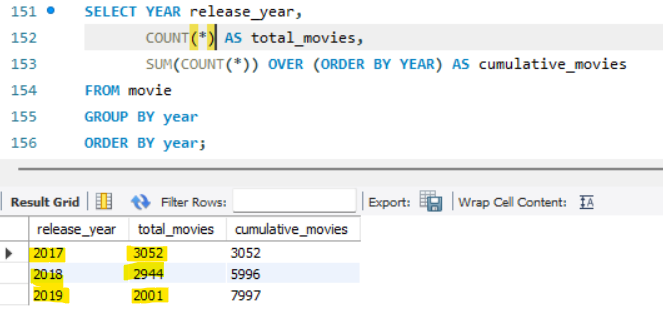


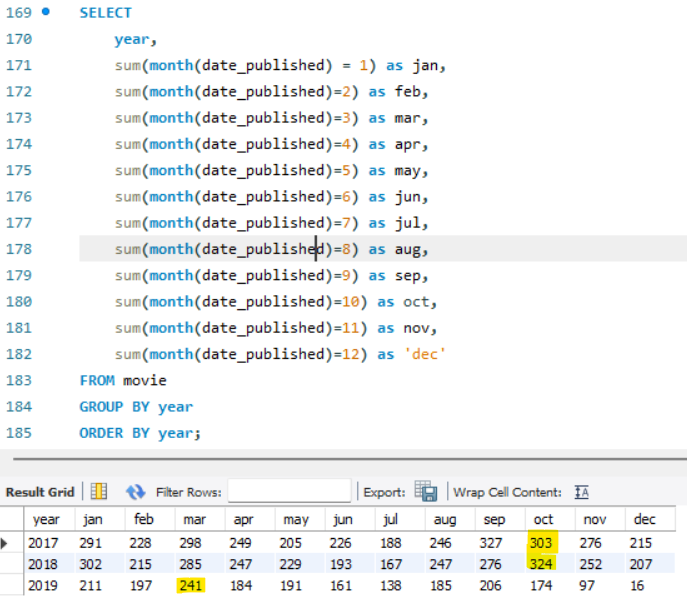
1. Determine the total number of movies released each year, and analyze how the trend changes month-wise.

This is the query to fetch the movie count for each year with a month-wise total in a row-wise format.

The SUM aggregate function is used to calculate the total number of movies released in each month. Additionally, the MONTH() function is applied to the date\_published column to categorize movies based on their release month. Finally, the data is grouped by year to present a structured view of movie releases over time.

* The dataset shows a declining trend in the number of movies released from 2017 to 2019.
* The peak months for movie releases were September and October, with consistently high numbers across years.
* December 2019 saw the least movie releases (only 16), possibly due to industry slowdowns or external factors.
* September 2017 had the highest releases (327), suggesting it might be a peak release period.

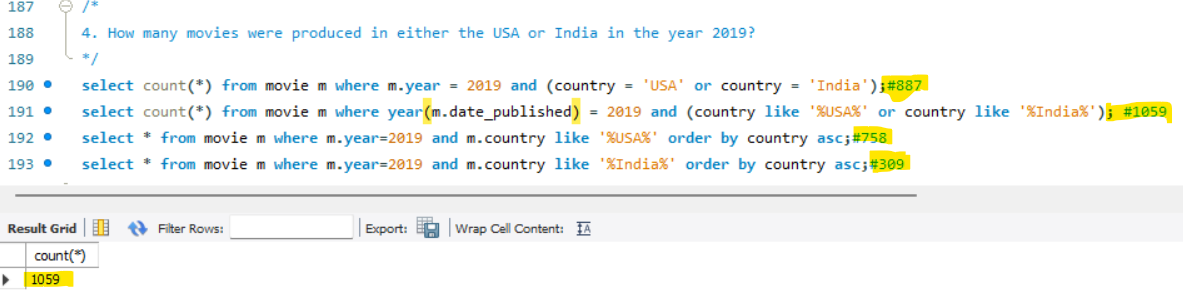




1. How many movies were produced in either the USA or India in the year 2019?

To retrieve movies produced in the year 2019 and released in USA or India, we need to use the WHERE condition along with the LIKE operator to filter records where the country column contains either USA or India.

The query below returns 1,059 movies that were released in 2019 in either USA or India.

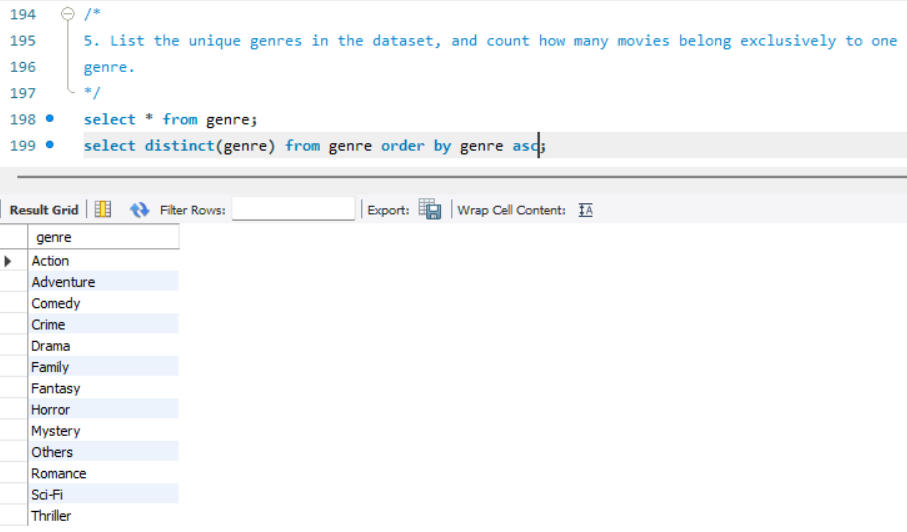


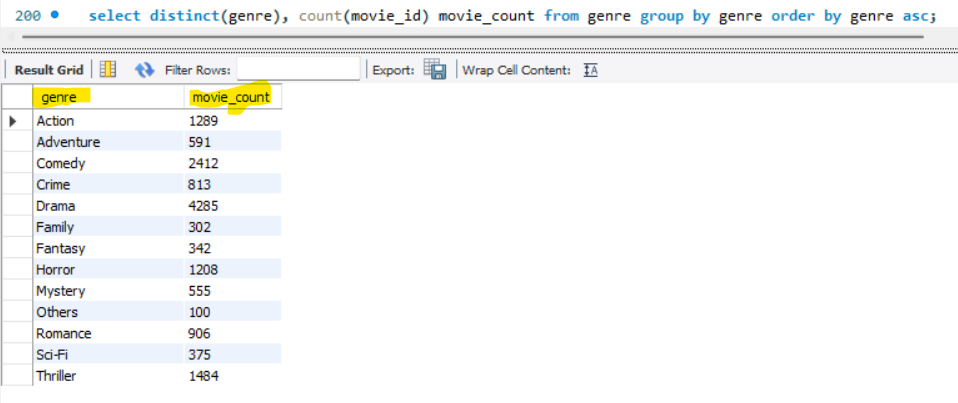
1. List the unique genres in the dataset, and count how many movies belong exclusively to one genre.

To get the unique genres, we need to use the DISTINCT function to retrieve unique genre types from the genre table.

Additionally, the COUNT function is used to calculate the number of movies in each genre.

The result below shows that Drama is the most common genre, with 4,285 movies, compared to other genres.



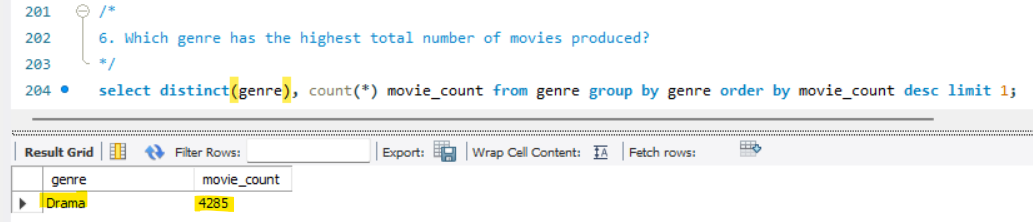


1. Which genre has the highest total number of movies produced?

In the previous query, we retrieved the total number of movies for each genre. To determine the genre with the highest number of movies,

we need to sort the movie count in descending order using the ORDER BY clause. Additionally, we use the LIMIT clause to restrict the result to only the top row.

The query returns Drama as the genre with the highest number of movies, totaling 4,285.



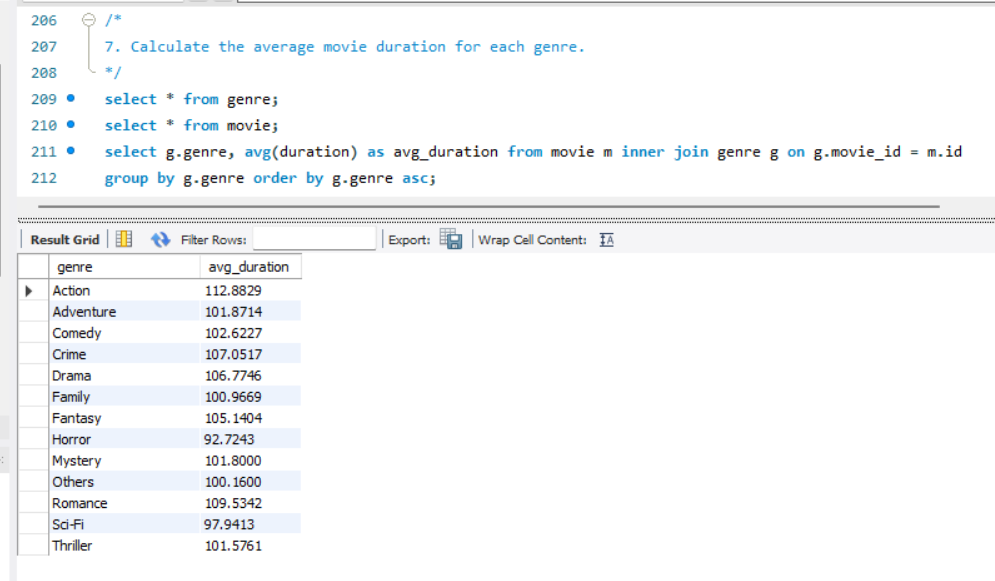
1. Calculate the average movie duration for each genre.

The query calculates the average duration of movies for each genre by joining the movie and genre tables using the movie\_id column.

It uses the AVG() function to compute the average duration and groups the results by genre. The results are sorted in ascending order of genre names.

* Action movies have the highest average duration (112.88 minutes), making them one of the longest genres.
* Romance movies also tend to be long, with an average duration of 109.53 minutes.
* Horror movies have the shortest average duration (92.72 minutes), suggesting they are typically more concise.
* Most other genres, including Comedy, Crime, Drama, and Thriller, fall within the 100-110 minute range.

This analysis helps in understanding the general length of movies based on their genre, which can be useful for filmmakers, streaming platforms, and audiences when selecting movies.

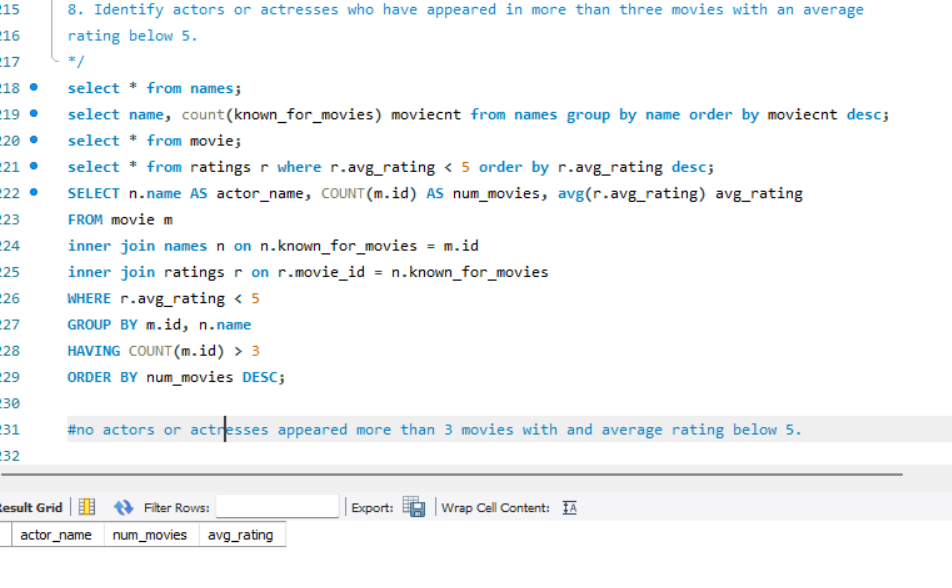


1. Identify actors or actresses who have appeared in more than three movies with an average rating below 5.

The query attempts to identify actors or actresses who have appeared in more than three movies where the average rating is below 5.

* No actors or actresses have appeared in more than three movies where the average rating is below 5.
* This suggests that actors generally do not consistently work on poorly rated films or that the dataset does not have such cases.

This insight helps in analyzing actor performance trends and movie quality associations.



1. Find the minimum and maximum values for each column in the ratings table, excluding the movie\_id column.

The query retrieves the minimum and maximum values for each numeric column in the ratings table, excluding the movie\_id column.

min(avg\_rating), max(avg\_rating):

Finds the lowest and highest average movie ratings.

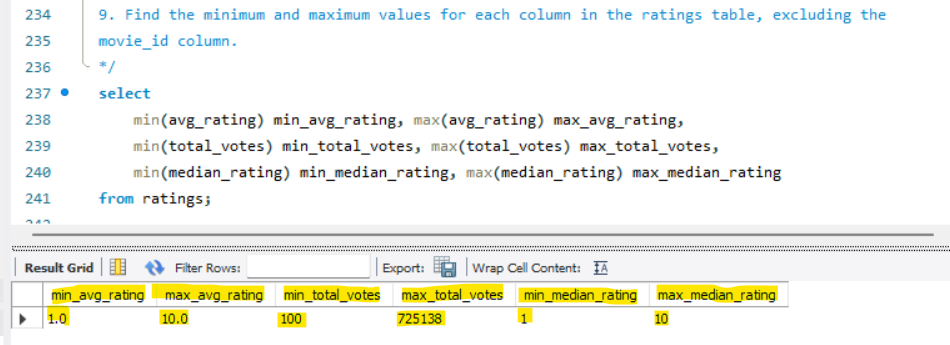
min(total\_votes), max(total\_votes):

Identifies the least and most number of votes received by any movie.

min(median\_rating), max(median\_rating):

Determines the minimum and maximum median ratings.

* Some movies received the lowest possible rating (1.0), while others achieved a perfect score (10.0).
* The total votes range from 100 (minimum engagement) to 725,138 (highly rated/popular movies).
* The median rating distribution also spans from 1 to 10, indicating a diverse mix of critically poor and highly acclaimed films.
* This analysis helps in understanding movie popularity trends and rating distributions across the dataset.



1. Which are the top 10 movies based on their average rating?

This analysis helps in identifying highly acclaimed movies based on viewer ratings.

The query retrieves the top 10 movies sorted by their average rating in descending order.

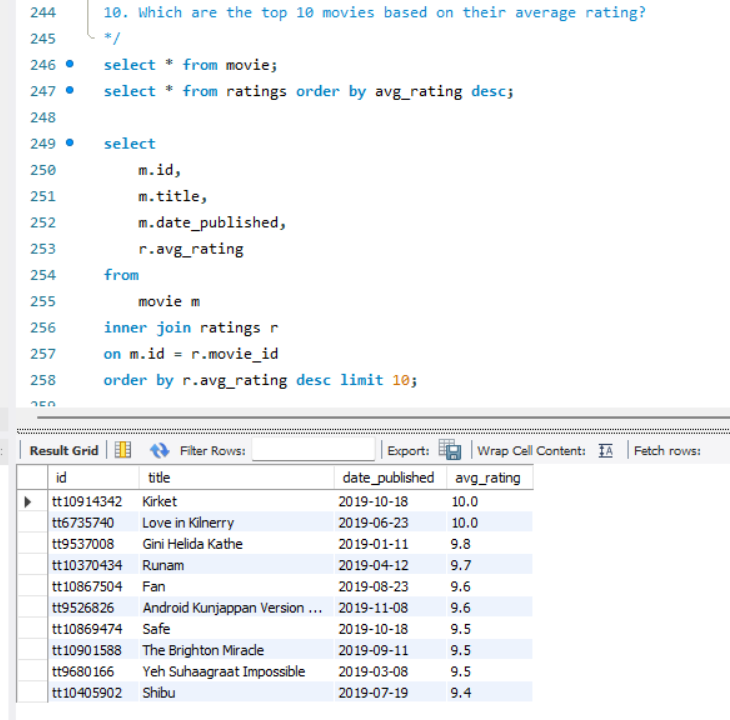
Retrieves movie details (id, title, date\_published).

Joins the movie table with the ratings table using movie\_id.

Orders the results based on avg\_rating in descending order.

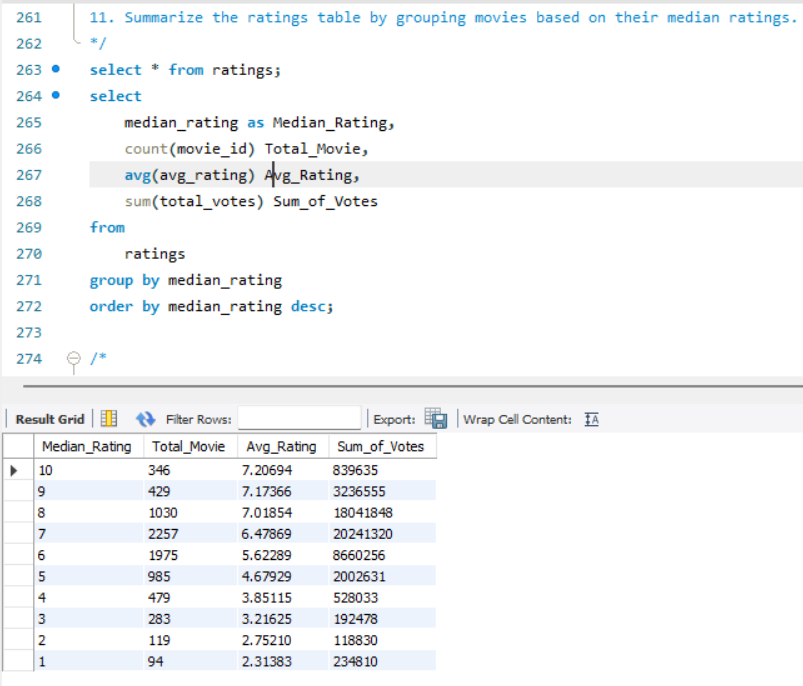
Limits the output to 10 movies (highest-rated).

* Two movies ("Kirket" and "Love in Kilnerry") received perfect ratings (10.0).
* Most top-rated movies were released in 2019, indicating strong audience engagement that year.
* The Indian movie "Android Kunjappan Version 5.25" (Malayalam) made it to the top with 9.6 rating.



1. Summarize the ratings table by grouping movies based on their median ratings.

* High-rated movies (9-10) are fewer in number but generally have higher votes per movie.
* Most movies cluster around the 6-7 range, showing that the majority are perceived as average or slightly above average.
* Lower-rated movies (1-3) exist but are relatively fewer, suggesting that truly bad movies are rare or receive fewer votes.



1. How many movies, released in March 2017 in the USA within a specific genre, had more than 1,000 votes?

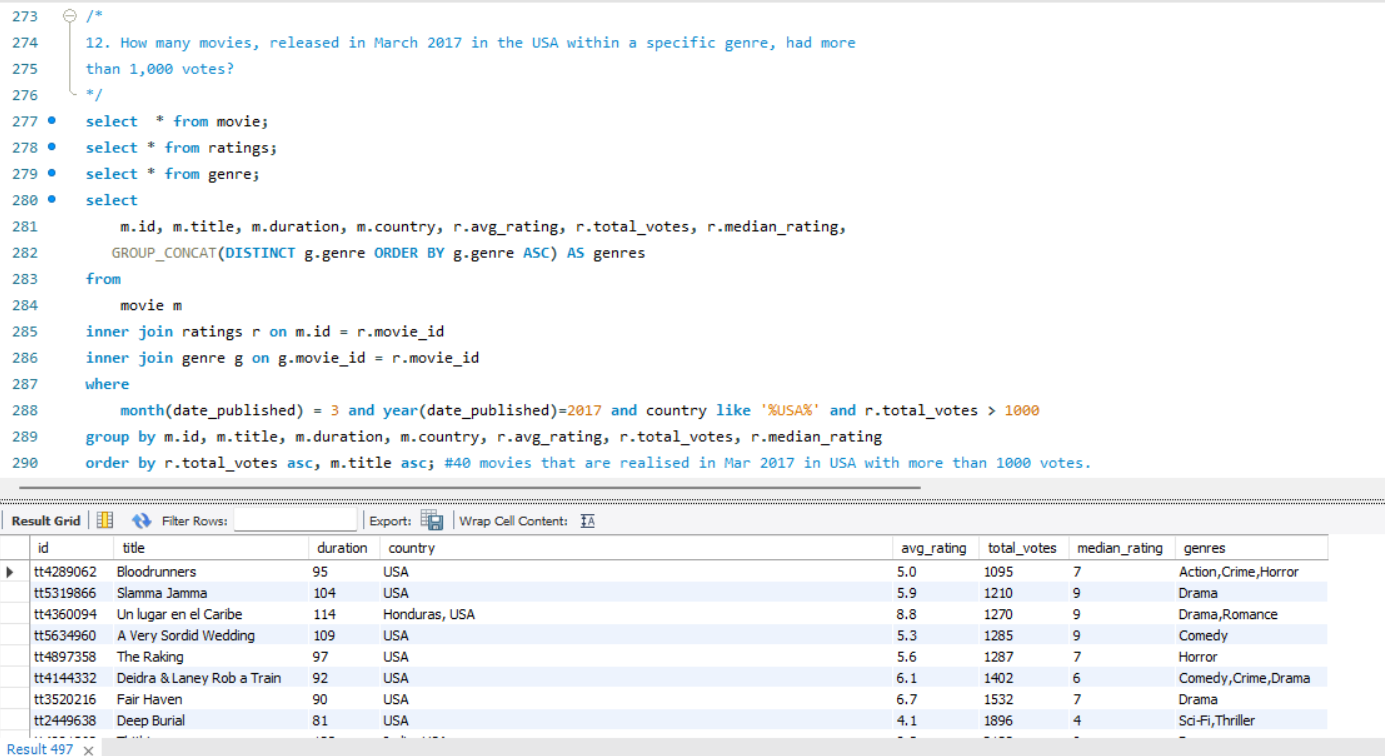
A total of 40 movies were released in March 2017 in the USA.

In the movie table, we only have the date\_published column. Therefore, we used the MONTH() function to filter movies released in March

and the YEAR() function to filter those from 2017.

Since the country column contains multiple values (e.g., South Africa, USA, Germany), we used the LIKE operator to fetch movies

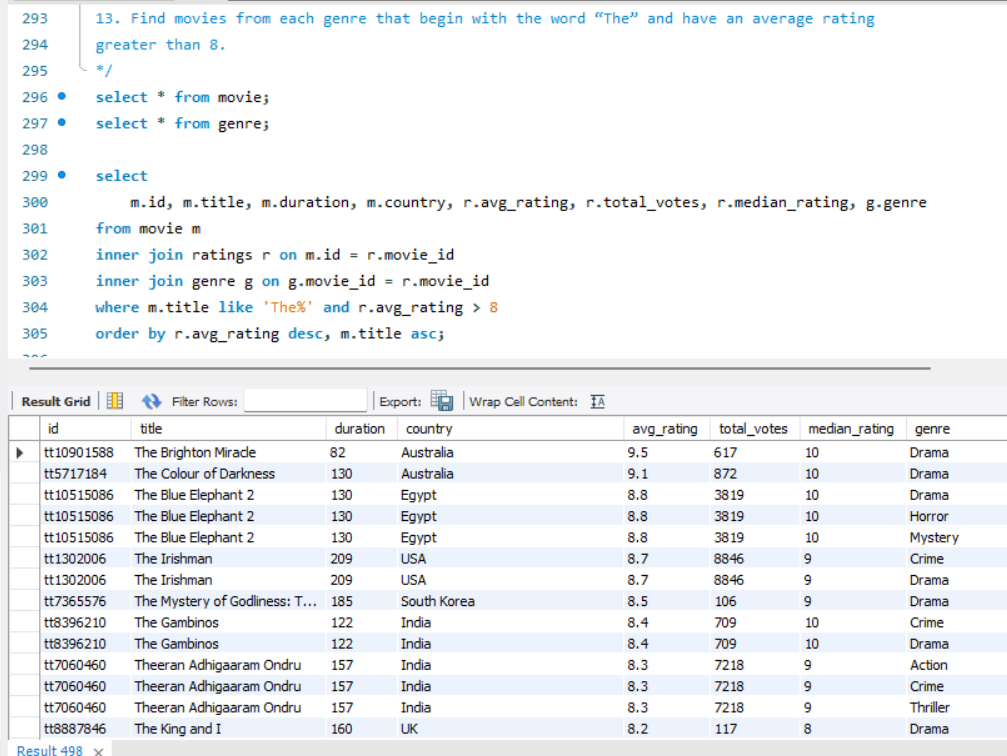
that include 'USA' as one of the countries.



1. Find movies from each genre that begin with the word “The” and have an average rating greater than 8.

Using the LIKE operator in the WHERE clause, we filtered genres that begin with the word 'The' and have an average rating above 8.

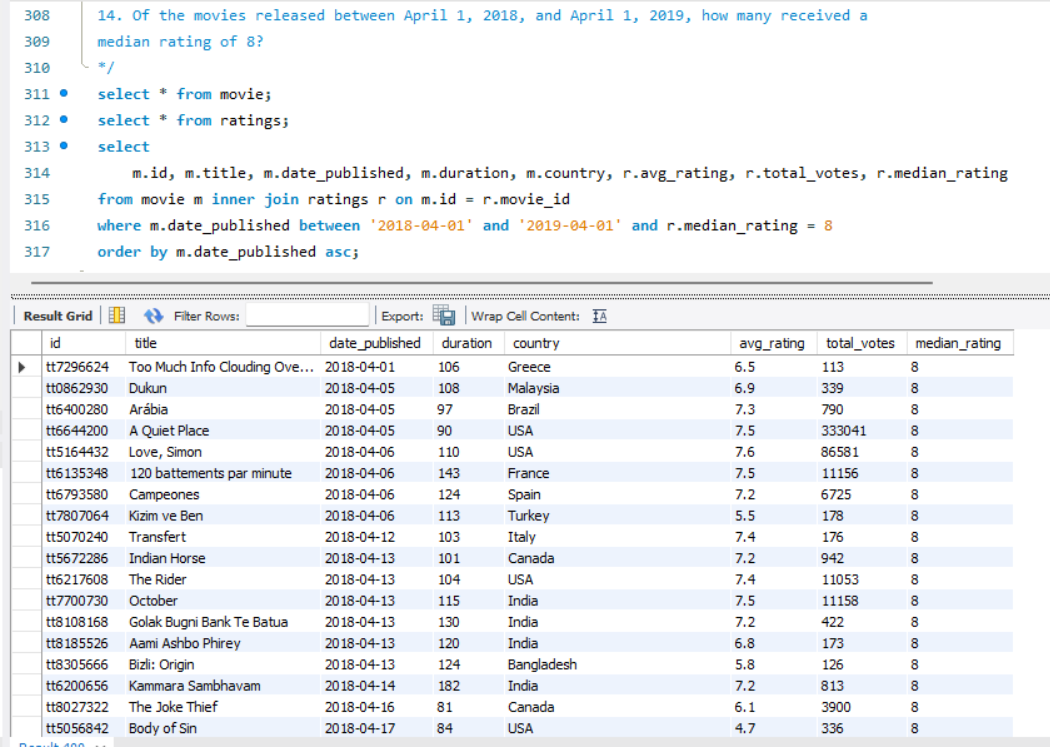
A total of 15 genres meet these criteria.



1. Of the movies released between April 1, 2018, and April 1, 2019, how many received a median rating of 8?

A total of 361 movies were released between 1st April 2018 and 1st April 2019 and received a median rating of 8.

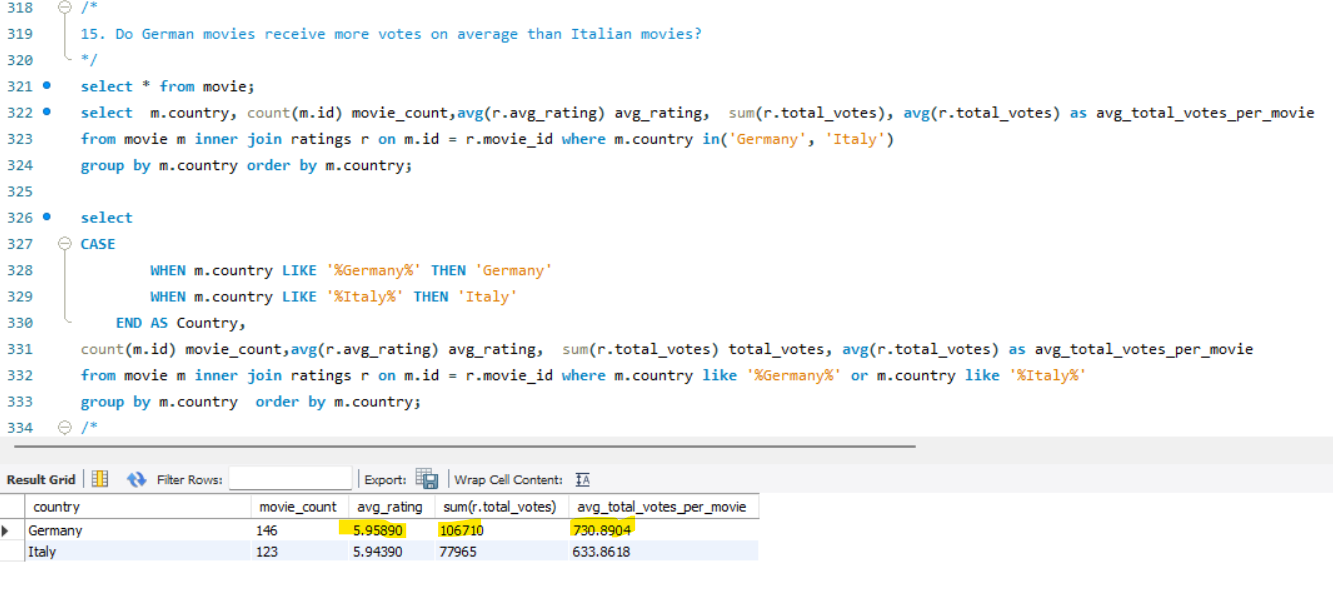
To achieve this, the BETWEEN keyword was used to filter movies within the specified date range (1st April 2018 to 1st April 2019) and with a median\_rating of 8.



1. Do German movies receive more votes on average than Italian movies?

German movies receive more votes on average (730.89 votes per movie) than Italian movies (633.86 votes per movie).

Both countries have similar average ratings (Germany: 5.96, Italy: 5.94), indicating no significant difference in viewer ratings.



1. Identify the columns in the names table that contain null values.

In the names table, there are five columns: id, name, height, date\_of\_birth, and known\_for\_movies.

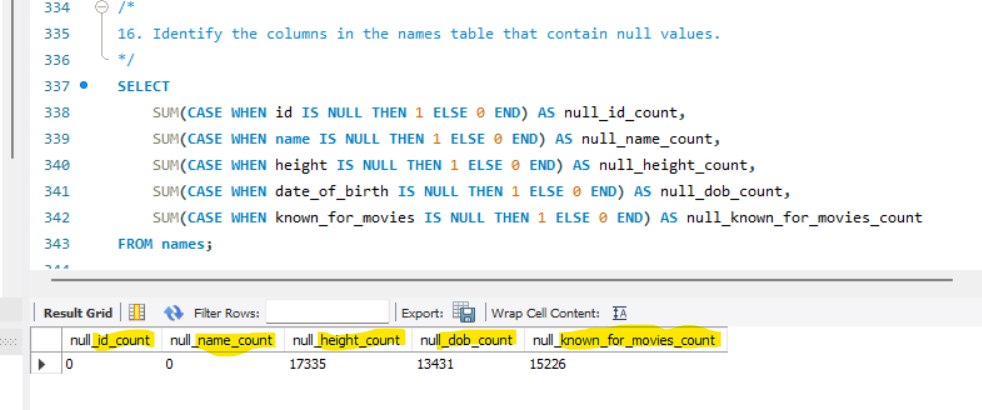
The id and name columns do not contain any NULL values.

The remaining columns contain NULL values as follows:

height → 17,335 NULL values

date\_of\_birth → 13,431 NULL values

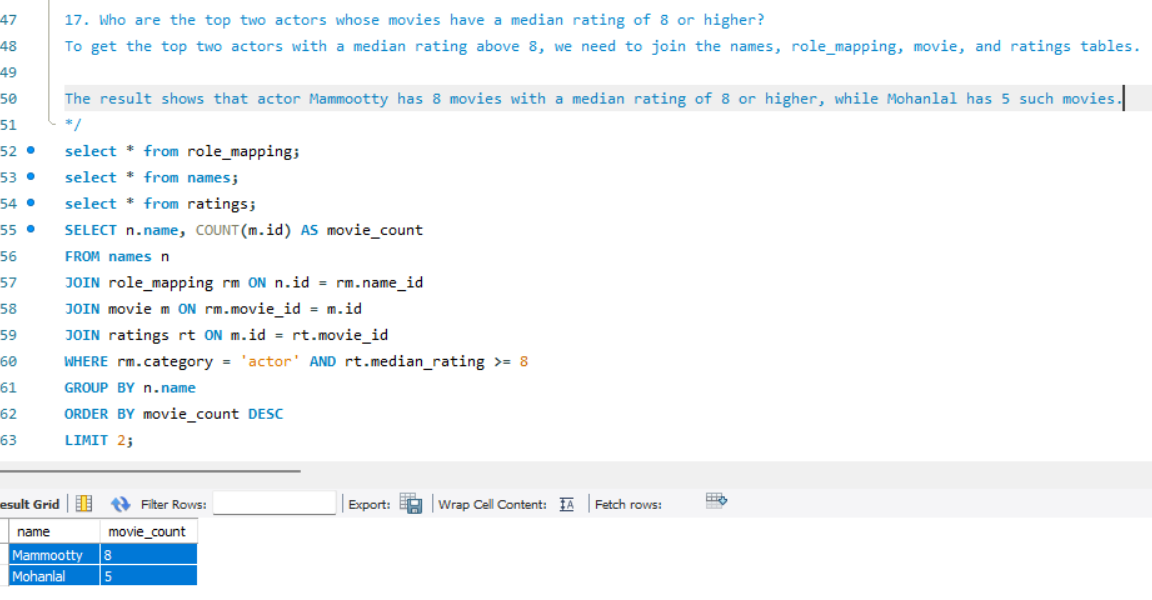
known\_for\_movies → 15,226 NULL values



1. Who are the top two actors whose movies have a median rating of 8 or higher?

To get the top two actors with a median rating above 8, we need to join the **names, role\_mapping, movie, and ratings** tables.

The result shows that **actor Mammootty has 8 movies** with a **median rating of 8 or higher**, while **Mohanlal has 5 such movies**.

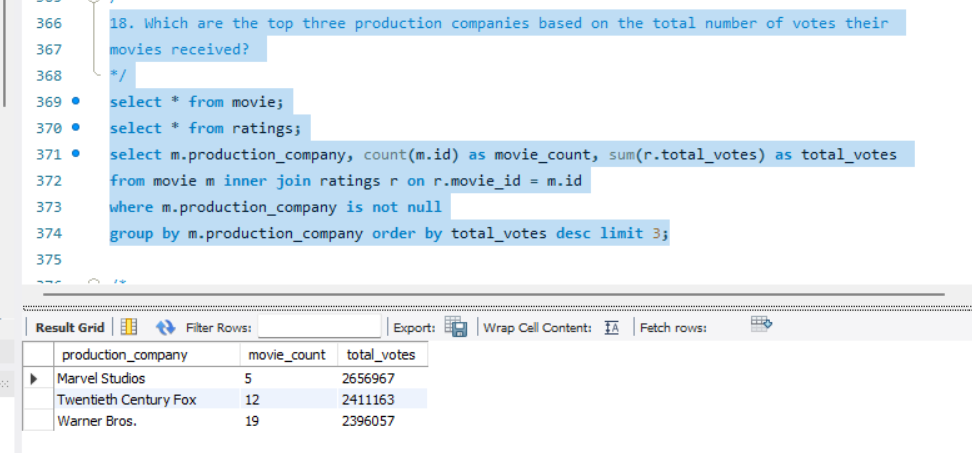


1. Which are the top three production companies based on the total number of votes their movies received?

Marvel Studios got the highest votes (2.65M) with just 5 movies.

Twentieth Century Fox (12 movies) and Warner Bros. (19 movies) had fewer votes per movie.

Marvel movies attract more audience engagement per film.



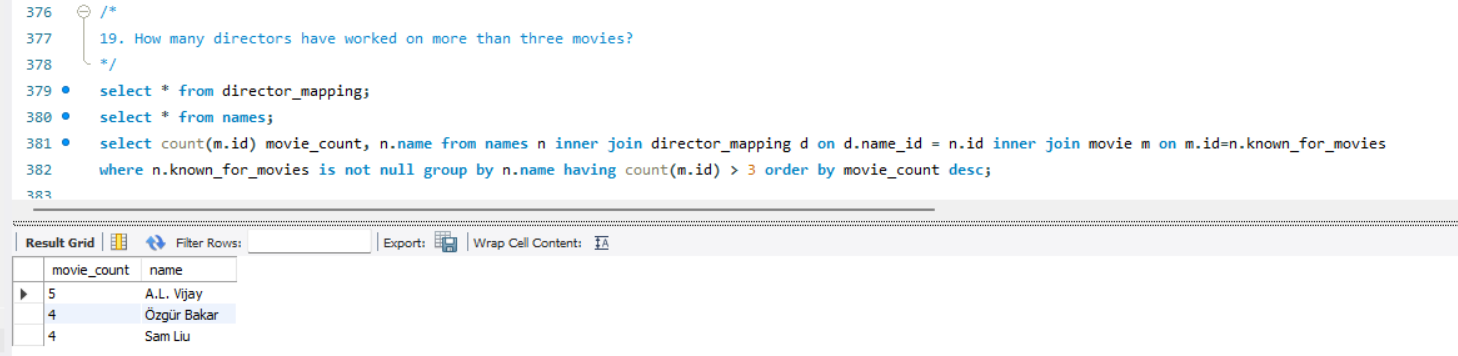
1. How many directors have worked on more than three movies?

Three directors have worked on more than three movies.

A.L. Vijay directed the most (5 movies).

Özgür Bakar and Sam Liu each directed 4 movies.

These directors are consistently involved in multiple projects.

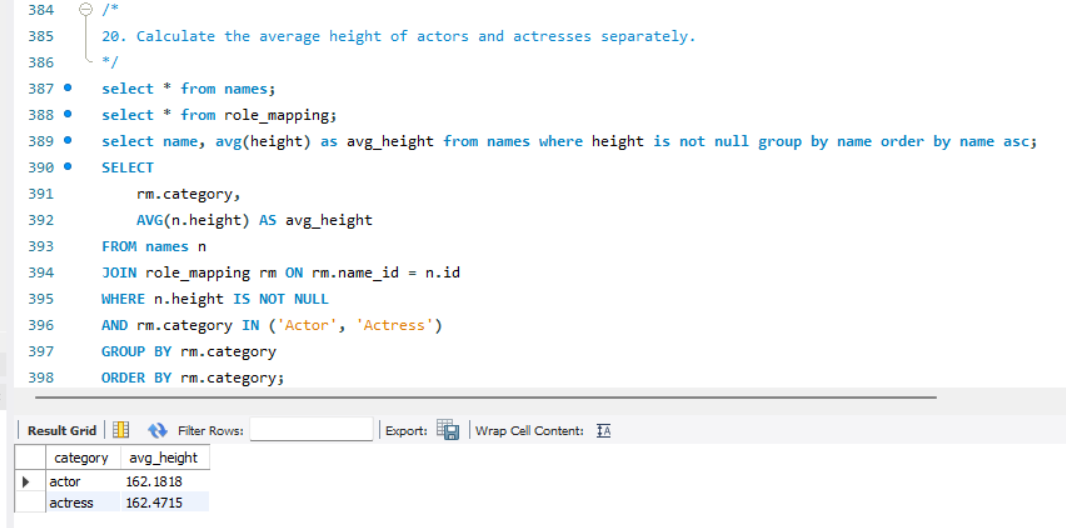


1. Calculate the average height of actors and actresses separately.

The average height of actors is 162.18 cm.

The average height of actresses is 162.47 cm.

There is only a 0.29 cm difference, meaning actors and actresses have nearly the same average height.

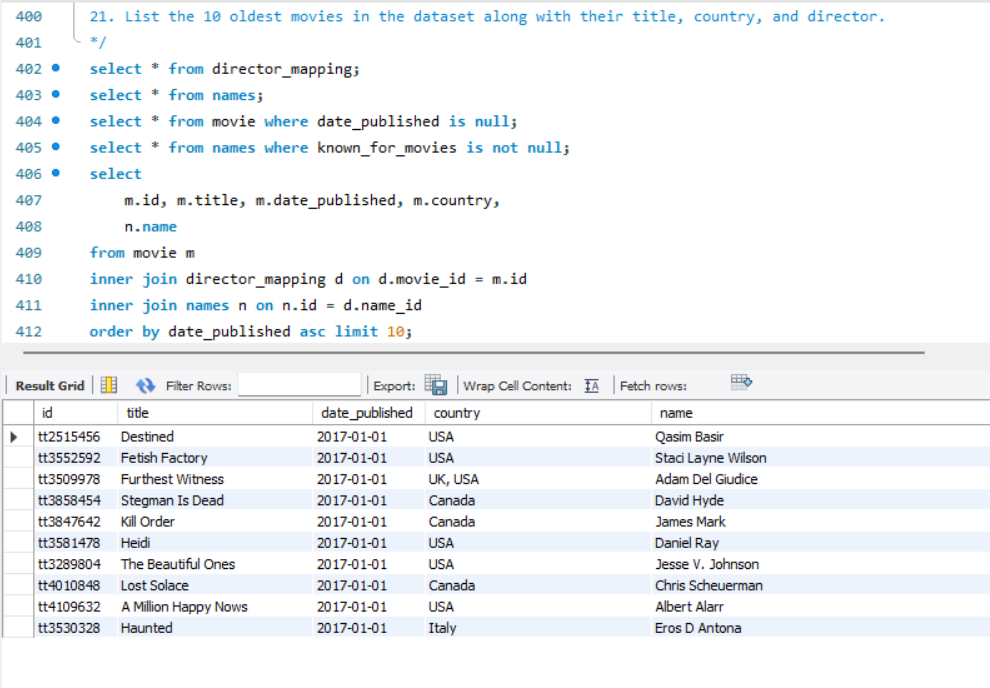


1. List the 10 oldest movies in the dataset along with their title, country, and director.

All 10 oldest movies in the dataset were released on January 1, 2017.

The movies come from different countries, including USA, Canada, UK, and Italy.

USA has the most movies on the list, followed by Canada.



1. List the top 5 movies with the highest total votes, along with their genres.

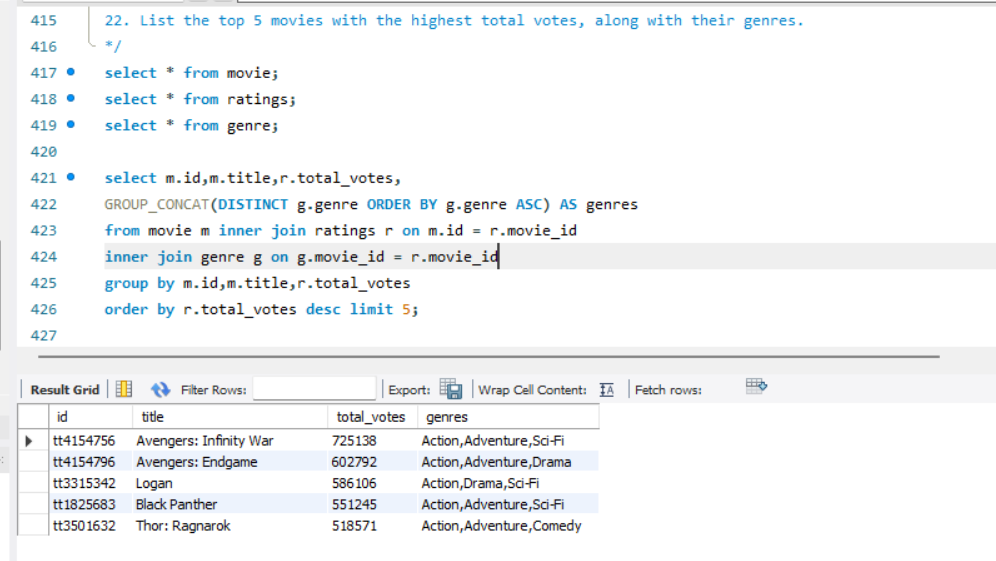
To fetch the top 5 movies with the highest total votes along with their genres, we need to join the movie table with ratings and genre tables.

Since the genre table contains multiple genre values for each movie, we use the GROUP\_CONCAT aggregate function to concatenate

multiple genre values into a single string.

We group the results by movie ID, title, and total votes, then order them by total votes in descending order and use LIMIT to fetch the top 5 rows.

* The most common genres among these movies are Action, Adventure, and Sci-Fi, with some featuring Drama or Comedy.
* Avengers: Infinity War received the highest number of total votes (725,138), followed by Avengers: Endgame (602,792).

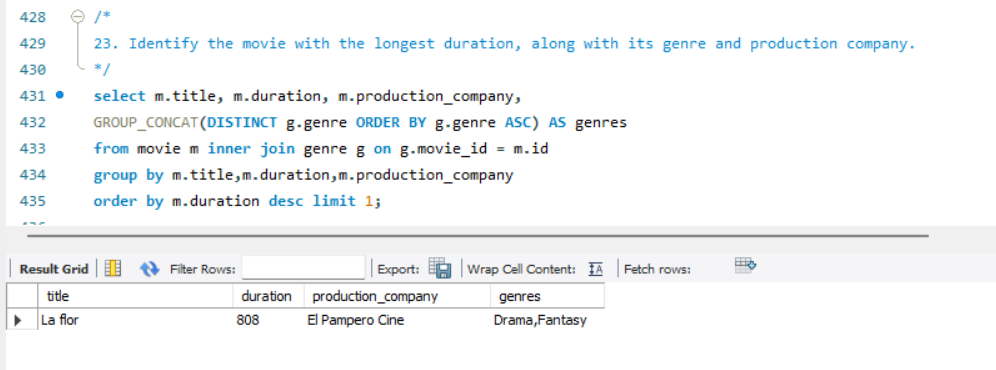


1. Identify the movie with the longest duration, along with its genre and production company.

To identify the movie with the longest duration, along with its genre and production company, we can use two approaches:

Sorting the duration in descending order and selecting the top row using LIMIT 1.

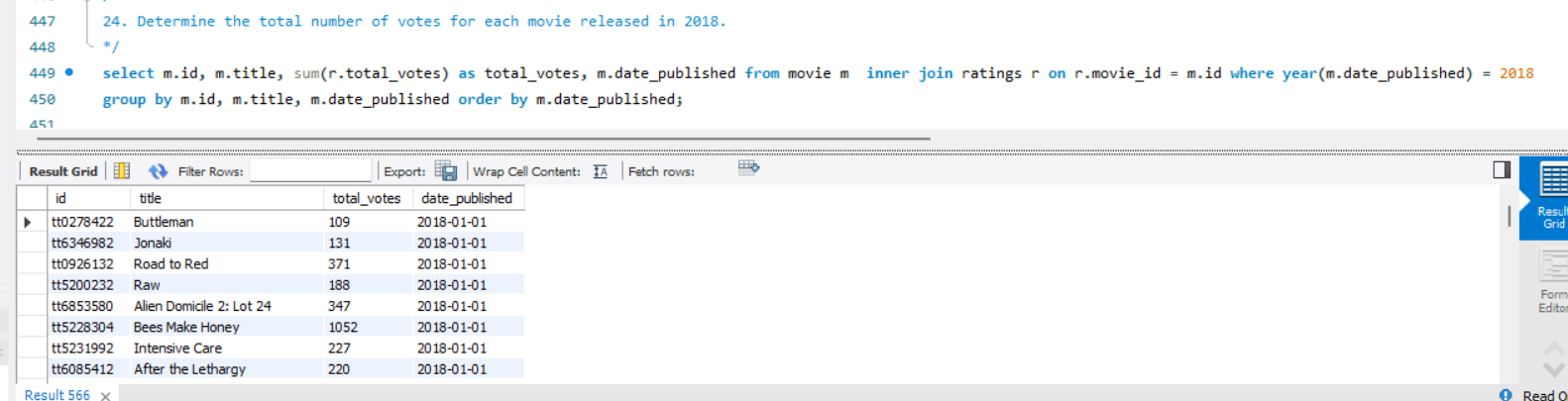
Using a WHERE condition to filter the movie with the maximum duration.



1. Determine the total number of votes for each movie released in 2018.

To get the total number of movies released in 2018 along with their total votes, we need to join the movie table with ratings,

filter for movies released in 2018, and group the data by movie\_id, title, and date\_published to calculate the total votes for each movie.



1. What is the most common language in which movies were produced?

To determine the most common language in which movies were produced, we need to analyze the languages column in the movie table.

Since this column may contain multiple languages in a single row (e.g., "English, French, Russian"), we should split these values and count each language separately.

The most common language for movie production is English, with 4,134 movies. This is followed by French (612), Spanish (566), Hindi (341), and German (335).

To determine the most common language:

We split the languages column into individual languages and stored them in a temporary table (temp\_movie\_languages).

Then, we counted the number of movies produced in each language.

The result was sorted in descending order to find the most frequently used language.

